

ROTATION SAMPLING FOR FUNCTIONAL DATA

David Degras

DePaul University

Abstract: Survey sampling methods provide cost-effective solutions for monitoring levels in large populations. Rotation designs, which replace a fraction of the sample at each survey occasion, yield better performances than fixed panels for this purpose. Motivated by recent applications in sensor networks, we extend rotation sampling to functional data (that is, continuous signals). We propose survey designs that replace the sample at fixed times via Markov chains. We then study the related Horvitz-Thompson estimator of the population mean function and quantify its large-sample behavior. For a given sample size, the mean estimation error can be decreased by suitably allocating the sample during replacements; the error variance can be dramatically reduced by increasing either the frequency or the intensity of replacements. We also extend the composite estimation method to the continuous-time setup. The new estimator combines the Horvitz-Thompson estimator, which only uses current data, with the estimated population change between an arbitrary past time and the current observation time. A simulation study shows that the proposed rotation designs and composite estimator considerably improve upon fixed panels and the Horvitz-Thompson estimator.

Key words and phrases: Design-based survey; functional data; rotation sampling; Horvitz-Thompson estimator; asymptotic theory; composite estimator.

1. Introduction

In various industrial, environmental and medical applications, sensor networks generate large volumes of data in continuous time. Due to cost or energy constraints, such collections of functional data (that is, curve data) often cannot be entirely observed. For instance, large electric utilities need to monitor the total consumption of their clients in order to adjust the power generation to the system load, predict future consumption, and determine pricing policies. These utilities however cannot read all of their client smart meters at each instant, as the network cannot process such large data transmission and the cost for data storage would be prohibitive. Under such observation constraints, survey sam-

pling methods offer competitive solutions for monitoring global parameters; see e.g., Chiky, Cubillé, Dessertaine, Hébrail, and Picard (2008) for a comparison between survey sampling and signal compression approaches.

Repeated surveys are well studied in the statistical literature. In the classical modeling framework where population parameters are fixed and only the sample selection is random, authors like Patterson (1950), Eckler (1955), Rao and Graham (1964), and Wolter (1979) have developed the theory of rotation sampling, a popular technique that consists in replacing a fraction of the sample with new population units at each survey occasion. They have established that different replacement strategies should be used depending on the survey's goal. For example, partial replacements of the sample are advantageous for assessing population levels while fixed panels are preferable for estimating level changes. Jointly with the study of sampling designs, they investigated estimation procedures such as minimum variance unbiased estimators and composite estimators, aiming to strike a balance between statistical and computational efficiency.

While traditional repeated surveys take place in discrete time and present important limitations on the sampling design and data collection, sensor networks such as smart electrical grids or audience measurement systems provide data every few seconds or minutes (essentially, in continuous time). This recent type of application enables sophisticated sampling designs and alleviates burdens of traditional surveys such as measurement errors, nonresponse and sample attrition. At the same time, it poses new statistical challenges related to the functional data/continuous-time setup. For example, Cardot and Josserand (2011) and Cardot, Degras, and Josserand (2012) extend the well-known Horvitz-Thompson [HT] estimator to continuous-time surveys. Based on results from Degras (2011) they address the construction of simultaneous confidence bands, which of course only makes sense in view of a continuous data-generating process. Studying the model-assisted survey of functional data, Cardot, Chaouch, Goga, and Labruère (2010) use functional principal component analysis and multivariate nonparametric regression to relate scalar auxiliary information to functional observations. Note that the above survey methods are all based on fixed panels.

In this paper we devise rotation designs for the survey of functional data. In contrast with traditional rotation designs which specify the sample before the

survey (see e.g., Eckler, 1955), we build rotation samples using Markov chains. As in Tikkiwal and Gupta (1991) this approach introduces a new randomization at each sample replacement. Importantly, it enables informative sampling: past measurements can be used to “improve” the sample at the next replacement by suitably updating the inclusion/exclusion probabilities. Examples of adaptive sampling rules would be to retain units in the sample with probabilities proportional to their current levels, or to optimize the sample allocation in stratified sampling based on the current strata sizes and variances.

In a second time, we examine the estimation of the population mean function based on our proposed rotation designs. Specifically, we quantify the large-sample behavior of the HT estimator in terms of the sampling rate, the replacement rate (that is, the fraction of the sample discarded at each replacement), and the number of replacements. We focus on the integrated squared error [ISE] which is a relevant measure for assessing a sampling strategy over a time interval. Our results show that when the sample size is constant over time, rotation samples and fixed panels yield the same mean ISE. On the other hand, the mean ISE can be sensibly reduced under rotation sampling by using stratification and suitably re-allocating the sample at replacement times. More importantly, the variance of the ISE can be decreased by several orders of magnitude when using rotation samples rather than fixed panels. This decrease is achieved by increasing either the frequency or the intensity of replacements.

Thirdly, this paper extends the composite estimation method (see e.g., Rao and Graham, 1964) to functional data. Unlike discrete-time surveys where composite estimators are updated with respect to the last survey occasion, there is no “last” occasion in continuous time. We thus define our composite estimator in reference to an arbitrary past observation time. This estimator is based on the HT estimator and an estimator of the population change. Note that the theory of minimum variance unbiased estimation (see e.g., Patterson, 1950) is of no help in our context as we do not consider infinite populations nor make assumptions on the temporal dependence. To finish, we present a numerical study with artificial electricity consumption data that confirms the superior performances of our rotation samples and composite estimator over fixed panels and the HT estimator.

The rest of the paper is organized as follows. The modeling framework and the HT estimator are presented in Section 2. The new rotation designs for functional data are defined in Section 3. The mean and covariance of the HT estimator are studied in Section 4. Section 5 contains our main result, namely the large-sample variance of the ISE under the studied rotation designs. The composite estimator is introduced in Section 6 and the simulation study is described in Section 7. Concluding remarks are offered in Section 8. The proofs of the main results are collected in the Appendix and other proofs are available online as supplementary material.

2. Statistical framework

Let $U_N = \{1, \dots, N\}$ be a finite population in which a deterministic curve $X_k(t), t \in [0, T]$, is associated to each unit $k \in U_N$. We study the estimation of the population mean function

$$\mu_N(t) = \frac{1}{N} \sum_{k \in U_N} X_k(t)$$

based on observations $X_k(t), k \in s(t)$, where $s(t) \subset U_N$ is a probability sample of fixed size $n(t)$. For mathematical simplicity, we assume perfect observations. If the sampled curves are observed at discrete times and/or with noise, interpolation or smoothing methods should be applied to the data. In this case the results of this paper will still hold under standard interpolation or nonparametric smoothing conditions. See for example Cardot and Josserand (2011) and Cardot, Degras and Josserand (2012). The sample $s(\cdot) = \{s(t), t \in [0, T]\}$ is selected according to a controlled probability distribution p_N over the function space $\mathcal{P}(U_N)^{[0, T]}$, where $\mathcal{P}(U_N)$ denotes the set of all subsets of U_N . For any units $k, l \in U_N$ and times $t, t' \in [0, T]$, the first and second order inclusion probabilities under p_N are written as $\pi_k(t) = \mathbb{P}(k \in s(t))$ and $\pi_{kl}(t, t') = \mathbb{P}(k \in s(t), l \in s(t'))$ respectively. Henceforth the subscript in U_N is dropped.

We use the estimator of Horvitz and Thompson (1952)

$$\hat{\mu}_{ht}(t) = \frac{1}{N} \sum_{k \in U} \frac{I_k(t)}{\pi_k(t)} X_k(t), \quad (2.1)$$

where $I_k(t)$ is the sample indicator function of $k \in U_N$ at time t : $I_k(t) = 1$ if $k \in s(t)$ and $I_k(t) = 0$ otherwise. The HT estimator is unbiased under p_N and

its covariance function is

$$\text{Cov}(\hat{\mu}_{ht}(t), \hat{\mu}_{ht}(t')) = \frac{1}{N^2} \sum_{k,l \in U} \frac{\Delta_{kl}(t, t')}{\pi_k(t)\pi_l(t')} X_k(t) X_l(t'),$$

where $\Delta_{kl}(t, t') = \text{Cov}(I_k(t), I_l(t')) = \pi_{kl}(t, t') - \pi_k(t)\pi_l(t')$.

The estimation accuracy can be improved by using a suitable stratification of the population. From now on we assume that U is partitioned into strata U_h of size N_h for $h = 1, \dots, H$. For each U_h , denote by $\mu_h(t) = (1/N_h) \sum_{k \in U_h} X_k(t)$ and $\gamma_h(t, t') = (1/(N_h - 1)) \sum_{k \in U_h} (X_k(t) - \mu_h(t))(X_k(t') - \mu_h(t'))$ the stratum mean and covariance functions. Let $n_h(t) = \#(s(t) \cap U_h)$ be the sample size in U_h at time t and $f_h(t) = n_h(t)/N_h$ be the sampling rate. If $s(t)$ is obtained by simple random sampling without replacement [SRSWOR] independently in each U_h , the HT estimator becomes

$$\hat{\mu}_{ht}(t) = \frac{1}{N} \sum_{h=1}^H \frac{1}{f_h(t)} \sum_{k \in U_h} I_k(t) X_k(t) \quad (2.2)$$

and its covariance rewrites as

$$\text{Cov}(\hat{\mu}_{ht}(t), \hat{\mu}_{ht}(t')) = \frac{1}{N^2} \sum_{h=1}^H \frac{1}{f_h(t)f_h(t')} \sum_{k,l \in U_h} \Delta_{kl}(t, t') X_k(t) X_l(t'). \quad (2.3)$$

3. Rotation designs for continuous-time surveys

The rotation sampling technique has been developed in the context of (discrete-time) traditional repeated surveys. In this section we extend rotation sampling to the continuous-time setup of functional data. We propose two novel probability schemes for selecting a time-varying sample $s(\cdot) = \{s(t), t \in [0, T]\}$ in a stratified population. These sampling designs, which we refer to as full replacement and partial replacement, share the following features:

- The strata samples $s_h(\cdot) = \{s(t) \cap U_h, t \in [0, T]\}$ are independent across strata.
- At time $\tau_0 = 0$, the strata samples $s_h(\tau_0)$ are obtained by SRSWOR.
- The $s_h(\cdot)$ can only be modified at fixed times $0 < \tau_1 < \dots < \tau_m < T$.

It remains to specify the evolution mechanisms of the discrete processes $\{s_h(\tau_r), r = 1, \dots, m\}$ under full and partial replacement.

1. **Full replacement.** In each stratum U_h , the successive samples $s_h(\tau_r)$, $r = 1, \dots, m$, are obtained by independent SRSWOR of $n_h(\tau_r)$ units in U_h .
2. **Partial replacement.** In each stratum U_h , a fraction $\alpha_h \in [0, 1]$ of $s_h(\cdot)$ is replaced at each time τ_r , $r = 1, \dots, m$. Specifically, $s_h(\tau_{r-1})$ is updated at time τ_r by the following independent operations:
 - discard $\alpha_h n_h(\tau_{r-1})$ units selected in $s_h(\tau_{r-1})$ by SRSWOR;
 - add $(n_h(\tau_r) - n_h(\tau_{r-1}) + \alpha_h n_h(\tau_{r-1}))$ units selected in $U_h \setminus s_h(\tau_{r-1})$ by SRSWOR.

Note that full replacement is not a special case of partial replacement with $\alpha_h = 1$. Indeed, for $r = 1, \dots, m$, $s_h(\tau_{r-1})$ and $s_h(\tau_r)$ are independent in the former case whereas they are disjoint (and thus dependent) in the latter case. In partial replacement we refer to the α_h as the replacement rates. For simplicity we assume that the α_h are constant over time and that the proposed sample replacements are possible without modifications, which entails that $\alpha_h n_h(\tau_{r-1}) \in \mathbb{N}$ and $n_h(\tau_{r-1}) \leq n_h(\tau_r) + \alpha_h n_h(\tau_{r-1}) \leq N_h$ for all h, r . Note that fixed panels are a special case of partial replacement with $\alpha_h = 0$ and $n_h(\cdot)$ constant over time.

We now determine the probability distribution of the sample $s(t)$ under the proposed sampling designs. The following result relies on an induction argument on the τ_r under partial replacement while it holds trivially under full replacement.

Proposition 1. *Consider either the full or the partial replacement design. For each stratum U_h and time $t \in [0, T]$, the probability distribution of the sample $s_h(t)$ is identical to the SRSWOR of $n_h(t)$ units in U_h .*

4. Covariance of the Horvitz-Thompson estimator

Here we derive the covariance function of the HT estimator (2.2) under the full and partial replacement designs of the previous section. Let $\nu(t)$ denote the number of sample replacements occurring before time t :

$$\nu(t) = \min \{r \in \{0, \dots, m\} : \tau_r \leq t\}$$

so that $t \in [\tau_{\nu(t)}, \tau_{\nu(t)+1})$ for all $t \in [0, T]$ with the convention $\tau_{m+1} = T$.

4.1. Covariance under the full replacement design

Under the full replacement design, $\hat{\mu}_{ht}(t)$ and $\hat{\mu}_{ht}(t')$ are independent if the sample has been replaced between times t and t' . Therefore the expression of the covariance (2.3) directly follows from the properties of SRSWOR. Let $\delta_{..}$ indicate the Krönecker delta.

Theorem 1. *Consider the full replacement design of Section 3. For all strata U_h , units $k, l \in U_h$, and times $t, t' \in [0, T]$,*

$$\Delta_{kl}(t, t') = (1 - f_h(t)) f_h(t) \delta_{\nu(t)\nu(t')} \frac{N_h \delta_{kl} - 1}{N_h - 1}.$$

As a consequence, the covariance of the HT estimator (2.2) is

$$\text{Cov}(\hat{\mu}_{ht}(t), \hat{\mu}_{ht}(t')) = \frac{1}{N} \sum_{h=1}^H \frac{N_h}{N} \frac{1 - f_h(t)}{f_h(t)} \gamma_h(t, t') \delta_{\nu(t)\nu(t')}.$$

This theorem will be commented in relation to partial replacement in the next section.

4.2. Covariance under the partial replacement design

In view of Proposition 1, it suffices to compute $\pi_{kl}(t, t') = \mathbb{E}(I_k(t)I_l(t'))$ for $k, l \in U_h$ ($h = 1, \dots, H$) and $0 \leq t \leq t' \leq T$ in order to determine $\Delta_{kl}(t, t')$ and (2.3). By definition of SRSWOR, the inclusion probabilities $\pi_{kk}(t, t')$ and $\pi_{kl}(t, t')$ do not depend on $k, l \in U_h$ ($k \neq l$). Since $\sum_k I_k(t) = n_h(t)$, one also sees that $\text{Cov}(\sum_{k \in U_h} I_k(t), \sum_{l \in U_h} I_l(t')) = N_h \Delta_{kk}(t, t') + N_h(N_h - 1) \Delta_{kl}(t, t') = 0$, where $k \neq l$ are two arbitrary units in U_h . Therefore, computing $\Delta_{kl}(t, t')$ amounts to computing $\Delta_{kk}(t, t')$ which in turn reduces to deriving $\mathbb{P}(k \in s_h(t') | k \in s_h(t))$.

The transition probabilities of the Markov chain $\{I_k(\tau_r), r = 0, \dots, m\}$ can be found by applying the Chapman-Kolmogorov equations. To that end, define

$$\lambda_h(t, t') = \prod_{r=\nu(t)+1}^{\nu(t')} \frac{1 - \alpha_h - f_h(\tau_r)}{1 - f_h(\tau_{r-1})} \quad (4.1)$$

for all times $0 \leq t \leq t' \leq T$ and extend $\lambda_h(t, t')$ as a symmetric function on $[0, T]^2$. Set $\lambda_h(t, t') = 1$ if $\nu(t) = \nu(t')$ and set to 1 all factors of $\lambda_h(t, t')$ for which $f_h(\tau_{r-1}) = 1$.

Lemma 1. *Consider the partial replacement design of Section 3. For all strata U_h , units $k \in U_h$, and times $0 \leq t \leq t' \leq T$,*

$$\begin{cases} \mathbb{P}(k \in s_h(t') | k \in s_h(t)) = (1 - f_h(t)) \lambda_h(t, t') + f_h(t'), \\ \mathbb{P}(k \in s_h(t') | k \notin s_h(t)) = f_h(t') - f_h(t) \lambda_h(t, t'). \end{cases}$$

In the proof of this lemma, the quantity $\lambda_h(t, t')$ turns out to be the product of the eigenvalues of the transition probability matrices of the previous Markov chain between times t and t' .

For any two real numbers x, y , write $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$. Exploiting Proposition 1 and Lemma 1, we obtain the covariance function (2.3) under partial replacement.

Theorem 2. *Consider the partial replacement design of Section 3. For all strata U_h , units $k, l \in U_h$, and times $t, t' \in [0, T]$,*

$$\Delta_{kl}(t, t') = (1 - f_h(t \wedge t')) f_h(t \wedge t') \frac{N_h \delta_{kl} - 1}{N_h - 1} \lambda_h(t, t').$$

As a consequence, the covariance of the HT estimator (2.2) is

$$\text{Cov}(\hat{\mu}_{ht}(t), \hat{\mu}_{ht}(t')) = \frac{1}{N} \sum_{h=1}^H \frac{N_h}{N} \frac{1 - f_h(t \wedge t')}{f_h(t \vee t')} \gamma_h(t, t') \lambda_h(t, t').$$

To provide insights into Theorems 1-2, we consider the situation where the sample sizes $n_h(\cdot)$ are constant over time. In this case the term $\lambda_h(t, t')$ in Theorem 2 simplifies into $(1 - \alpha_h / (1 - f_h))^{| \nu(t) - \nu(t') |}$. We can then compare the estimator covariance under the full and partial replacement designs, examining the “no replacement” strategy (that is, a fixed panel) as a special instance of partial replacement ($\alpha_h = 0$). On the diagonal blocks $[\tau_r, \tau_{r+1}]^2, r = 1, \dots, m$, $\text{Cov}(\hat{\mu}_{ht}(t), \hat{\mu}_{ht}(t'))$ is identical under the three previous types of replacement. Outside these blocks, the estimator covariance is zero under full replacement whereas it is a weighted average of the covariance functions γ_h under a fixed panel. For partial replacement the covariance structure depends on the replacement rates α_h . When α_h increases in $[0, 1 - f_h]$ for some h , the estimator covariance between given times t, t' decreases (assuming positivity of γ_h). In the special case where $\alpha_h = 1 - f_h$ for all h , the covariance is the same as under full replacement. For values $\alpha_h > 1 - f_h$, the covariance becomes unstable in the sense that the

term $(1 - \alpha_h/(1 - f_h))^{\nu(t) - \nu(t')}$ changes sign on every block $[\tau_q, \tau_{q+1}] \times [\tau_r, \tau_{r+1}]$. For $\alpha_h \notin \{0, 1 - f_h\}$ and regularly spaced τ_r (that is, $|\nu(t) - \nu(t')| \leq Cm|t - t'|$ for some finite constant C), the estimator covariance decreases at an exponential rate as $|t - t'|$ increases.

4.3. Mean Integrated Squared Error

To measure the accuracy of an estimator $\hat{\mu}_N$ of μ_N over $[0, T]$, we use the integrated squared error

$$\text{ISE} = \int_0^T (\hat{\mu}_N(t) - \mu_N(t))^2 dt.$$

As seen in Sections 2-3, the HT estimator (2.2) is unbiased and, when the sample sizes $n_h(\cdot)$ are constant over time, its variance function is the same under the full and partial replacement designs (and in particular under fixed panels). In this case, the HT estimator has the same mean integrated squared error

$$\text{MISE} = \int_0^T \mathbb{E} (\hat{\mu}_N(t) - \mu_N(t))^2 dt$$

under both designs. The MISE can however be reduced (in comparison to fixed panels) by using the full or partial replacement designs with suitable time-varying sample sizes $n_h(\cdot)$. Specifically, the variance of $\hat{\mu}_{ht}(t)$ can be minimized at each replacement time τ_r by choosing $n_h(\tau_r)$ according to the classical Neyman allocation rule $n_h(\tau_r) = c_r N_h \sqrt{\gamma_h(\tau_r, \tau_r)}$ with the constant c_r such that $\sum_h n_h(\tau_r) = n(\tau_r)$. See for example Fuller [2009, p. 21] for more details.

5. Asymptotic results for the ISE

We now determine the variance of the ISE for the HT estimator (2.2) based on the full or partial replacement designs of Section 3. Starting with the general HT estimator (2.1), this variance expresses as

$$\begin{aligned} \text{Var}(\text{ISE}) &= \iint_{[0, T]^2} \text{Cov} \left(\{\hat{\mu}_{ht}(t) - \mu_N(t)\}^2, \{\hat{\mu}_{ht}(t') - \mu_N(t')\}^2 \right) dt dt' \\ &= \frac{1}{N^4} \iint_{[0, T]^2} \sum_{i, j, k, l \in U} \frac{X_i(t) X_j(t) X_k(t') X_l(t')}{\pi_i(t) \pi_j(t) \pi_k(t') \pi_l(t')} \Delta_{ijkl}(t, t') dt dt', \end{aligned} \quad (5.1)$$

where we have introduced the four-fold cross-covariance function

$$\Delta_{ijkl}(t, t') = \text{Cov}(\{I_i(t) - \pi_i(t)\} \{I_j(t) - \pi_j(t)\}, \{I_k(t') - \pi_k(t')\} \{I_l(t') - \pi_l(t')\}).$$

While (5.1) can be computed exactly if the sample sizes $n_h(\cdot)$ are constant over time, large-sample approximations are required under time-varying $n_h(\cdot)$. The asymptotic framework of such approximations is detailed in Section 5.1, intermediate results are presented in Section 5.2, and the main results are given in Section 5.3.

5.1. Asymptotic framework

We let the strata sizes N_h , sample sizes $n_h(\cdot)$, replacement rates α_h ($h = 1, \dots, H$), and number of replacements m depend on the population size N and go to infinity (except for α_h) together with N . The number H of strata and the observation period $[0, T]$ stay fixed. We make the following assumptions.

- (A1) The curves X_k , $k \geq 1$, are integrable and uniformly bounded on $[0, T]$.
- (A2) $\int_0^{\tau_r} g(t)dt = r/(m+1)$ for $r = 1, \dots, m$, where g is a continuous density function supported by $[0, T]$.
- (A3) The sampling rate functions $f_h(\cdot)$ stay bounded away from zero as $N \rightarrow \infty$.
- (A4) The covariance functions γ_h converge uniformly on $[0, T]^2$ as $N \rightarrow \infty$ and have continuous limits.
- (A5) The number of replacements is dominated by the strata sizes: $m = o(\min_h(N_h))$ as $N \rightarrow \infty$.

The number H of strata, although fixed, can be large. The condition $N_h \rightarrow \infty$ is not restrictive as, typically, small strata U_h are fully observed and do not contribute to the estimation error. In (A1) the individual curves X_k are allowed to have discontinuity jumps. However (A4) requires that the strata covariance functions can be uniformly approximated by continuous functions, which entails that at any time t , at most an asymptotically negligible fraction of the $X_k(t)$ may have discontinuity jumps. This assumption is needed under full replacement to approximate $\gamma_h(t, t')$ around the diagonal $\{t = t'\}$ by the variance $\gamma_h(t, t)$. Assumption (A2) ensures that the replacement times are regularly spaced, while (A3) is necessary for the consistent estimation of μ_N . Finally (A5) is needed under partial replacement to approximate the transition probabilities in and out of the sample for fixed population units.

5.2. Intermediate results

Here we derive the cross-covariance function $\Delta_{ijkl}(t, t')$. Recall that the time-varying samples $s_h(\cdot)$ are independent across strata. By expressing the population mean function μ_N as the weighted average $\sum_h (N_h/N) \mu_h$ of the strata mean functions and using the corresponding decomposition for the HT estimator (2.2), it follows that

$$\text{Var (ISE)} = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \iint_{[0, T]^2} \sum_{i, j, k, l \in U_h} \frac{\Delta_{ijkl}(t, t')}{n_h^2(t) n_h^2(t')} X_i(t) X_j(t) X_k(t') X_l(t') dt dt'. \quad (5.2)$$

Writing $\tilde{X}_k(t) = X_k(t) - \mu_h(t)$ for $k \in U_h$ and $h = 1, \dots, H$, simple algebra shows that

$$\begin{aligned} & \sum_{i, j, k, l \in U_h} \Delta_{ijkl}(t, t') X_i(t) X_j(t) X_k(t') X_l(t') \\ &= \sum_{i, j, k, l \in U_h} \mathbb{E} (I_i(t) I_j(t) I_k(t') I_l(t')) \tilde{X}_i(t) \tilde{X}_j(t) \tilde{X}_k(t') \tilde{X}_l(t') \\ & \quad - N_h^2 f_h(t) f_h(t') (1 - f_h(t)) (1 - f_h(t')) \gamma_h(t, t) \gamma_h(t', t'). \end{aligned} \quad (5.3)$$

The sum in the right-hand side of (5.3) can be developed using the properties of SRSWOR. We denote the asymptotic equivalence of two real sequences (a_N) and (b_N) by $a_N \sim b_N$.

Proposition 2. *Consider either the full or the partial replacement design of Section 3. Assume (A1). Let $\{i^*, j^*, k^*, l^*\}$ be four distinct units in a given stratum U_h . Then*

$$\begin{aligned} & \sum_{i, j, k, l \in U_h} \mathbb{E} (I_i(t) I_j(t) I_k(t') I_l(t')) \tilde{X}_i(t) \tilde{X}_j(t) \tilde{X}_k(t') \tilde{X}_l(t') \\ & \sim (C_1(t, t') \gamma_h(t, t) \gamma_h(t', t') + C_2(t, t') \gamma_h^2(t, t')) N_h^2 \end{aligned}$$

uniformly in $t, t' \in [0, T]$ as $N_h \rightarrow \infty$, where

$$\begin{aligned} C_1(t, t') &= \mathbb{E} (I_{i^*}(t) I_{k^*}(t')) - \mathbb{E} (I_{i^*}(t) I_{j^*}(t) I_{k^*}(t')) - \mathbb{E} (I_{i^*}(t) I_{k^*}(t') I_{l^*}(t')) \\ & \quad + \mathbb{E} (I_{i^*}(t) I_{j^*}(t) I_{k^*}(t') I_{l^*}(t')) \end{aligned}$$

and

$$\begin{aligned} C_2(t, t') &= 2 \mathbb{E} (I_{i^*}(t) I_{i^*}(t') I_{k^*}(t) I_{k^*}(t')) - 4 \mathbb{E} (I_{i^*}(t) I_{i^*}(t') I_{j^*}(t) I_{k^*}(t')) \\ & \quad + 2 \mathbb{E} (I_{i^*}(t) I_{j^*}(t) I_{k^*}(t') I_{l^*}(t')). \end{aligned}$$

Under the full replacement design, the previous functions C_1 and C_2 can be expressed in terms of $f_h(t)$ and $f_h(t')$ and $\text{Var}(\text{ISE})$ can readily be computed. For partial replacement however, additional results are required.

We now provide a result on the distribution of the sample in a given subset conditional on past values of the sample. Denote conditional distributions by $\mathcal{L}(\cdot|\cdot)$. Fix a stratum U_h and a subset $D \subset U_h$. The Markovian nature of $\{s_h(\tau_r), r = 0, \dots, m\}$ and the properties of SRSWOR (namely, the probability that the sample contains D only depends on the size of D) yield the following result.

Lemma 2. *Consider the partial replacement design of Section 3. Then for all $0 \leq t \leq t' \leq T$,*

$$\mathcal{L}(s_h(t') \cap D \mid s_h(t)) = \mathcal{L}(s_h(t') \cap D \mid s_h(t) \cap D).$$

The lemma states that at time t , all information relative to the future distribution of $s_h(t') \cap D$ is contained in $s_h(t) \cap D$.

Using the Chapman-Kolmogorov equations, large-sample approximations and linear algebra, we derive the transition probabilities in and out of the sample for two units of a given stratum.

Proposition 3. *Consider the partial replacement design of Section 3 and assume (A3)-(A5). For all units $k, l \in U_h$ ($k \neq l$) and times $0 \leq t \leq t' \leq T$, it holds as $N \rightarrow \infty$ that*

$$\begin{cases} \mathbb{P}(k, l \in s_h(t') \mid k, l \in s_h(t)) \sim [(1 - f_h(t))\lambda_h(t, t') + f_h(t')]^2, \\ \mathbb{P}(k, l \in s_h(t') \mid k \in s_h(t), l \notin s_h(t)) \\ \quad \sim [-f_h(t)(1 - f_h(t))\lambda_h^2(t, t') + f_h(t')(1 - 2f_h(t))\lambda_h(t) + f_h^2(t')], \\ \mathbb{P}(k, l \in s_h(t') \mid k, l \notin s_h(t)) \sim [(1 - f_h(t))\lambda_h(t, t') - (1 - f_h(t'))]^2. \end{cases}$$

5.4. Variance of the Integrated Squared Error

Based on the previous findings, we can now state the main results.

Theorem 3. *Consider the HT estimator (2.2) based on the full replacement design of Section 3. Assume (A1)-(A2)-(A3)-(A4). Then as $N \rightarrow \infty$,*

$$\text{Var}(\text{ISE}) \sim \frac{2}{mN^2} \int_0^T \left(\sum_{h=1}^H \frac{N_h}{N} \frac{1 - f_h(t)}{f_h(t)} g(t) \gamma_h(t, t) \right)^2 dt.$$

Theorem 4. *Consider the HT estimator (2.2) based on the partial replacement design of Section 3. Assume (A1)-(A2)-(A3)-(A5). Then as $N \rightarrow \infty$,*

$$\text{Var (ISE)} \sim \frac{2}{N^2} \iint_{[0,T]^2} \left(\sum_{h=1}^H \frac{N_h}{N} \frac{1 - f_h(t)}{f_h(t')} \lambda_h(t, t') \gamma_h(t, t') \right)^2 dt dt'.$$

Under additional assumptions, it is possible to find the asymptotic expression of $\lambda_h(t, t')$ in the previous theorem. Let G be an antiderivative of the density g in (A2).

Corollary 1. *Assume the conditions of Theorem 4 and (i) the sample sizes $n_h(\cdot)$ are constant over time, (ii) $\lim_{N \rightarrow \infty} (\alpha_h m / (1 - f_h)) = c_h < \infty$ exists. Then as $N \rightarrow \infty$,*

$$\text{Var (ISE)} \sim \frac{2}{N^2} \iint_{[0,T]^2} \left(\sum_{h=1}^H \frac{N_h}{N} \frac{1 - f_h}{f_h} \exp(-c_h |G(t) - G(t')|) \gamma_h(t, t') \right)^2 dt dt'.$$

The previous condition (ii) is reasonable since $(\alpha_h m)/T$ is the average sample replacement rate per unit time, which in practice stays bounded.

Under the assumptions of Theorems 3-4 and Corollary 1, we now compare the full and partial replacement designs in terms of variability of the ISE. As in Section 4 we include fixed panels as a special case of partial replacement where $\alpha_h = c_h = 0$. In comparison to fixed panels, partial replacement with $c_h > 0$ induces an exponentially decreasing function in $\text{Var}(\text{ISE})$. The decrease is all the larger as the c_h are large and the data are (positively) correlated. In comparison to partial sample replacements, the full replacement design divides the order of $\text{Var}(\text{ISE})$ by a factor m , which massively stabilizes the estimation performance.

Remark 1. *If the survey objectives include the integral $I_N = \int_0^T \mu_N(t) dt$, then $\hat{I}_N = \int_0^T \hat{\mu}_{ht}(t) dt$ provides an unbiased estimator whose variance follows from the previous results. As before, in comparison to fixed panels, partial replacement of the sample reduces $\text{Var}(\hat{I}_N)$ by an exponentially decreasing function and full replacement divides the order of $\text{Var}(\hat{I}_N)$ by a factor m .*

6. Composite estimation

The estimation of $\mu_N(t)$ can be improved by using past data in addition to the current observations $X_k(t), k \in s(t)$. Following the principle of composite

estimation (see e.g., Eckler, 1955), we utilize the partial replacement design of Section 3 and define a new estimator $\hat{\mu}_c(t)$ in an iterative fashion as a linear combination of the Horvitz-Thompson estimator $\hat{\mu}_{ht}(t)$, which is only based on the current observations, and of the estimator $\hat{\mu}_c(t - \delta)$ updated with the estimated change in μ_N between $t - \delta$ and t , where $\delta \in (0, T)$ is a time window to be specified. Let $0 \leq t \leq t' \leq T$. If $|\nu(t) - \nu(t')| \leq 1$, the estimator

$$\begin{aligned} \hat{\Delta}\mu_N(t, t') &= \frac{1}{N} \sum_{k \in U} \frac{I_k(t)I_k(t')}{\pi_{kk}(t, t')} (X_k(t') - X_k(t)) \\ &= \sum_{h=1}^H \frac{N_h}{N} \left(\frac{1}{(1 - \alpha_h) n_h(t)} \sum_{k \in s_h(t) \cap s_h(t')} (X_k(t') - X_k(t)) \right) \end{aligned} \quad (6.1)$$

is unbiased for the population mean change $\Delta\mu_N(t, t') = \mu_N(t') - \mu_N(t)$. If $|\nu(t) - \nu(t')| > 1$, the previous estimator is extended as

$$\hat{\Delta}\mu_N(t, t') = \hat{\Delta}\mu_N(t, \tau_{\nu(t)+1}) + \sum_{r=\nu(t)+1}^{\nu(t')-1} \hat{\Delta}\mu_N(\tau_r, \tau_{r+1}) + \hat{\Delta}\mu_N(\tau_{\nu(t')}, t'). \quad (6.2)$$

The composite estimator is then defined by $\hat{\mu}_c(t) = \hat{\mu}_{ht}(t)$ for $t \in [0, \tau_1)$ and

$$\hat{\mu}_c(t) = Q \hat{\mu}_{ht}(t) + (1 - Q) (\hat{\mu}_c((t - \delta)_+) + \hat{\Delta}\mu_N((t - \delta)_+, t)) \quad (6.3)$$

for $t \in [\tau_1, T]$, where $Q \in [0, 1]$ must be specified and $x_+ = \max(x, 0)$. Note that if $\alpha_h = 0$ with constant sample size $n_h(\cdot)$ for all h , or if $\delta = 0$, or if $Q = 1$, then $\hat{\mu}_c(t)$ reduces to $\hat{\mu}_{ht}(t)$. In this sense, the composite estimator can be thought of as a shrinkage estimator reverting to the Horvitz-Thompson estimator. The parameters α_h, δ , and Q govern the tradeoff between past and present data in the estimation. The theoretical study of the estimator (6.3) and the development of automated selection procedures for the α_h, δ and Q are beyond the scope of this paper. However, important insights for selecting these parameters are provided in the next section.

7. Numerical study

This section examines the numerical performances of the HT estimator (2.2) and composite estimator (6.3) based on the full and partial replacement designs of Section 3. The simulations are based on real electricity consumption curves studied by Cardot and Josserand (2011). In this study the electricity usage of

$N = 18902$ French firms was recorded every half hour during two weeks. Based on the functional principal components analysis (FPCA) of the second week of data, we generate realistic new data as follows:

$$X_k(t) = \mu_N^*(t) + \sum_{\ell=1}^3 Z_{\ell k} \phi_{\ell}(t) + \varepsilon_k(t), \quad (7.1)$$

where $k \in \{1, \dots, N\}$, $N = 10000$ and $t \in [0, 1]$. The term μ_N^* in (7.1) is the mean function of the real data; the $Z_{\ell k}$ are independent random variables distributed as $N(0, \sigma_{\ell}^2)$; $\sigma_1^2, \sigma_2^2, \sigma_3^2$ and ϕ_1, ϕ_2, ϕ_3 are the first eigenvalues and eigenfunctions of the FPCA (see Figure 7.1). The process ε_k is a Gaussian white noise; it is independent of the $Z_{\ell k}$ and verifies $\text{Var}(\varepsilon_k(t)) = \delta^2 \text{Var}(\sum_{\ell=1}^3 Z_{\ell k} \phi_{\ell}(t))$ with $\delta = 3\%$. The target function is $\mu_N = (1/N) \sum_{k=1}^N X_k$.

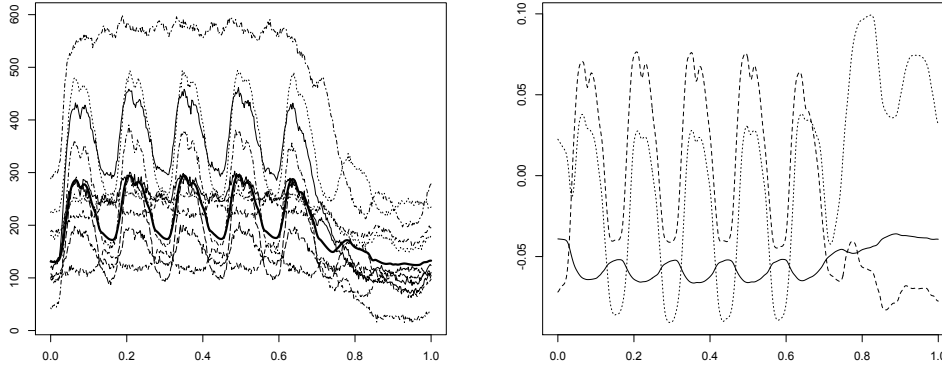


Figure 7.1: Typical curves X_k with mean function μ in thick solid line (left panel) and eigenfunctions ϕ_1, ϕ_2, ϕ_3 (right panel).

In the simulations the curves X_k are discretized at $d = 400$ equidistant points $t_j = j/(d + 1)$ and grouped in $H = 5$ strata according to their average levels $\int_0^1 X_k(t)dt$. The strata cutoffs are determined by the box plot. We set the sample size to $n(t) = 0.05 N$ at each instant $t \in [0, 1]$ and consider two sample allocations: [PROP] allocation proportional to stratum size with $n_h(t) = nN_h/N$ and [OPT] optimal allocation with $n_h(t) \propto N_h \gamma_h(t, t)^{1/2}$ (Neyman allocation). The strata sizes and sample sizes in each stratum are displayed in Table 7.1 (the

intervals in the bottom row are the range of $n_h(\cdot)$ under optimal allocation).

h	1	2	3	4	5
N_h	34	2466	5000	2456	44
PROP: n_h	2	123	250	123	2
OPT: $n_h(t)$	[1,2]	[126,134]	[226,240]	[129,138]	[1,2]

Table 7.1: Strata sizes and sample sizes under different allocations.

The replacement times are defined as $\tau_r = t_{3r}$, $r = 1, \dots, m$, with $m = [d/3] = 133$, which corresponds to a period of 1h15 between successive replacements. The sample $s(\cdot)$ is built either by full or partial replacement, with the replacement rates $\alpha_1 = \dots = \alpha_H := \alpha$ and $\alpha \in \{0, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1\}$ under partial replacement. Each type of replacement (full or partial) is combined with each type of allocation (proportional or optimal). For each combination we approximate the mean and variance of the ISE for the HT estimator (2.2) based on 5,000 to 100,000 Monte Carlo simulations of $s(\cdot)$.

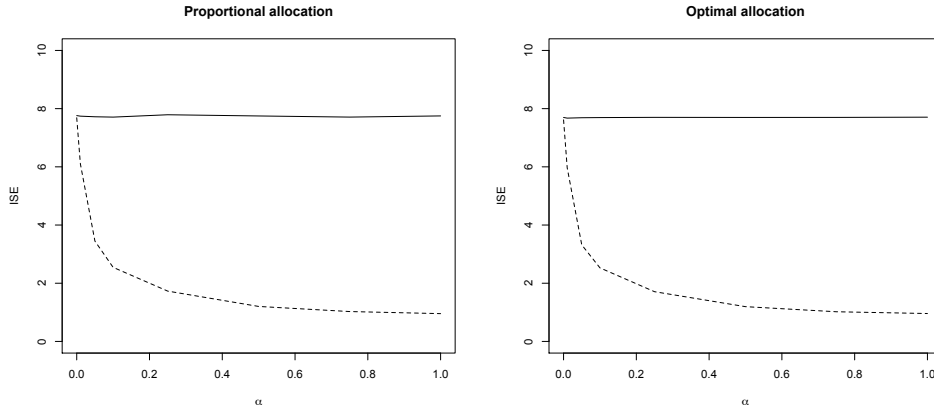


Figure 7.2: Horvitz-Thompson estimator of the mean function. Mean value (solid line) and standard deviation (dashed line) of the ISE in function of the replacement rate α .

In Table 7.1 the sample sizes under proportional and optimal allocation are quite close. This closeness stems from the data-generation mechanism (7.1) and the stratification. Indeed, in our simulations as in reality the main source of

variation for electricity usage is scale effects; in contrast longitudinal variations are qualitatively similar in the population. Also, the stratification used produces homogeneous strata with comparable standard deviations $\gamma_h(t, t)^{1/2}$ at a given time t . It is thus not surprising that in Figure 7.2 the estimation error is almost identical under proportional and optimal allocation. In the same figure, the MISE remains constant for all values of the replacement rate α as predicted by theory (see Section 4.3). A tiny reduction of the MISE can be achieved by using optimal allocation rather than proportional allocation; a much larger decrease would occur if the strata variances fluctuated differently over time. Importantly, Figure 7.2 illustrates the rapid decrease of $\text{Var}(\text{ISE})$ as α increases to 1 (see Corollary 1). In comparison to a fixed panel ($\alpha = 0$), both the partial replacement design with $\alpha = 1$ and the full replacement design reduce the standard deviation of the ISE by a factor 8.

In the second part of the simulations we turn our attention to the composite estimator (6.3). Keeping the previous data, population stratification, sample sizes $n_h(t)$ and replacement times τ_r , $r = 1, \dots, m$, we select the sample $s(\cdot)$ by partial replacement and proportional allocation. The replacement rate α varies in $\{i/10 : i = 1, \dots, 5\}$ and the tuning parameters of the estimator satisfy $\delta \in \{1.25, 2.5, 6, 12, 24, 48\}$ and $Q \in \{i/10 : i = 0, \dots, 10\}$. Note that because the strata samples $s_1(\cdot)$ and $s_5(\cdot)$ have size 2, $\alpha = 0.5$ is the largest value for which (6.3) can be computed. To facilitate the interpretation, δ is expressed in hours (recall that the survey duration is one week).

The mean estimation error MISE is displayed in Figure 7.3 in terms of α , δ and Q . To simplify the visualization, two parameters vary in each graph with the third one held fixed. The figure reveals a strong increase in the MISE as $Q \rightarrow 0$. This increase is very fast for small δ and slower as δ increases. For all δ , values of Q close to 1 (that is, when the composite estimator is close to the HT estimator) yield reasonable performances. On the other hand, Table 7.2 indicates that the optimal MISE is an increasing function of δ ($\delta = 25\text{mn}$ is the time step used for the discretization grid t_j , $j = 1, \dots, d$). This result is satisfactory in practice: it means that excellent performances can be obtained while retaining only very recent past data. This greatly alleviates data storage. Table 7.2 also shows that in order to obtain the best MISE, α must increase

and Q must decrease as δ decreases. A possible explanation for the two previous findings is that as the composite estimator $\hat{\mu}_c(t)$ gets updated with respect to an increasingly distant time point $t - \delta$, the estimation of the population change $\Delta\mu_N(t - \delta, t)$ becomes increasingly unreliable; to temper this, $\hat{\Delta\mu}_N(t - \delta, t)$ must be based on more matched units (i.e. smaller α) and more emphasis must be placed on the estimation based on the current data (i.e. larger Q). The above analysis results for the MISE also hold for the variance of the ISE.

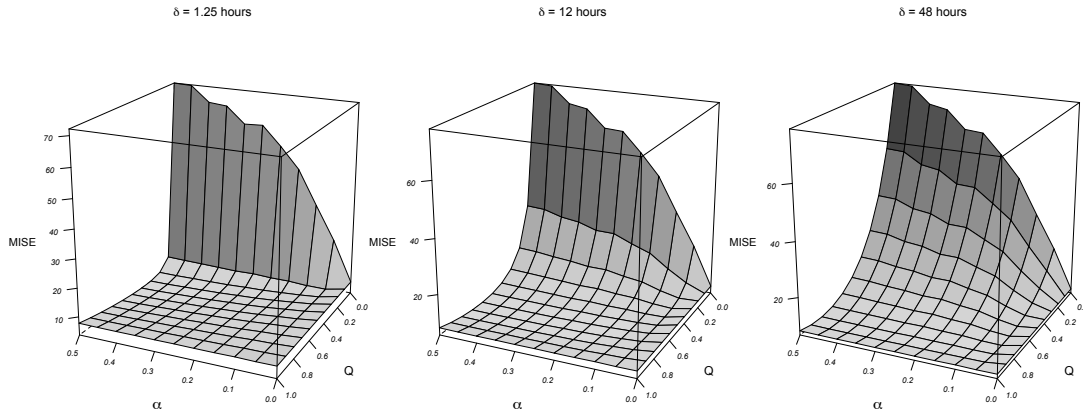


Figure 7.3: MISE of the composite estimator in terms of α and Q with δ fixed.

δ	25mn	50mn	1.25h	2.5h	6h	12h	24h	48h
$\min_{\alpha, Q}(\text{MISE})$	3.11	3.26	3.28	3.60	4.21	5.10	5.86	6.21
α_{min}	0.5	0.5	0.5	0.5	0.45	0.3	0.2	0.05
Q_{min}	0.1	0.2	0.2	0.4	0.5	0.6	0.8	0.7

Table 7.2: Optimal MISE and arguments α, Q of the composite estimator in terms of δ .

In simulations not included here, the results of this section have been seen to hold with other types of stratification (e.g., k -means) and other replacement frequencies m . Also, the full and partial replacement designs produce excellent performances for the estimation of the integral $\int_0^T \mu_N(t)dt$ (see Remark 1).

8. Discussion

With this paper we have taken initial steps in the exploration of time-varying samples with functional data. We have developed rotation sampling designs as well as a composite estimation method. Their good statistical properties have been established in theory and in practice. In particular these methods have been seen to outperform fixed panels and the Horvitz-Thompson estimator.

Hereafter we mention possible extensions of this work and related research directions. A first extension pertains to informative sampling. In addition to the adaptive rules mentioned in the introduction, an interesting possibility for the partial replacement design would be to let the replacement rates α_h be random and time-varying. It would then be sensible to adjust $\alpha_h(t)$ to the balance between transversal and longitudinal variations in the stratum U_h at time t : $\alpha_h(t)$ would be larger when transversal variations dominate longitudinal ones and smaller in the opposite case. A related problem is to find adaptive strategies that adjust the frequency of replacements at a given time (with our notations, the τ_r) in function of the magnitude of transversal variations (see e.g., Marbini, 2003). Of course, when developing a survey sampling strategy for a particular application, the search for statistical optimality should be balanced with the various costs and constraints incurred by sampling. Regarding the composite estimator introduced in the paper, it would be useful to study its theory (in particular derive its variance) and to devise automated procedures for selecting its tuning parameters δ and Q . Beyond composite estimation, effective survey estimation methods in the functional data setup could be based on functional linear models or nonparametric function regression approaches (see e.g., Ramsay and Silverman, 2005, and Ferraty and Vieu, 2006). Finally, our approach could be enhanced by incorporating auxiliary information in the survey. Model-assisted paradigms are available when the sampled curves are observed over the entire study period (Cardot, Chaouch, Goga, and Labruère, 2010) but methods for time-varying samples remain to be developed. In this regard it would be enlightening to compare the benefits of using auxiliary information in design-based approaches (for survey weights) versus model-assisted approaches.

Acknowledgment

The author acknowledges the Statistical and Applied Mathematical Sciences Institute (SAMSI) for its support during this research. He also thanks Professor Cardot (Université de Bourgogne) for motivating this project.

Appendix

Proof of Proposition 2

The sum under study can be decomposed as $\sum_{\ell=1}^4 A_\ell(t, t')$, where

$$A_\ell(t, t') = \sum_{\substack{i,j,k,l \in U_h \\ \mathcal{C}_{ijkl}=\ell}} \mathbb{E} (I_i(t)I_j(t)I_k(t')I_l(t')) \tilde{X}_i(t)\tilde{X}_j(t)\tilde{X}_k(t')\tilde{X}_l(t')$$

and $\mathcal{C}_{ijkl} = \#\{i, j, k, l\}$. To compute the A_ℓ , we derive $\mathbb{E} (I_i(t)I_j(t)I_k(t')I_l(t'))$ based on the properties of SRSWOR and develop sums $\sum \tilde{X}_i(t)\tilde{X}_j(t)\tilde{X}_k(t')\tilde{X}_l(t')$ using the identity $\sum_{k \in U_h} \tilde{X}_k(t) = 0$. Let i^*, j^*, k^*, l^* be four distinct units in U_h .

We begin with the straightforward calculation of $A_1(t, t')$:

$$A_1(t, t') = \mathbb{E} (I_{i^*}(t)I_{i^*}(t')) \sum_k \tilde{X}_k^2(t)\tilde{X}_k^2(t'). \quad (1)$$

The term $A_2(t, t')$ can be expressed as follows:

$$\begin{aligned} A_2(t, t') = \mathbb{E} (I_{i^*}(t)I_{k^*}(t')) & \left[(N_h - 1)^2 \gamma_h(t, t) \gamma_h(t', t') - \sum_{k \in U_h} \tilde{X}_k^2(t)\tilde{X}_k^2(t') \right] \\ & + 2 \mathbb{E} (I_{i^*}(t)I_{i^*}(t')I_{k^*}(t)I_{k^*}(t')) \left[(N_h - 1)^2 \gamma_h^2(t, t') - \sum_{k \in U_h} \tilde{X}_k^2(t)\tilde{X}_k^2(t') \right] \\ & - 2 \left[\mathbb{E} (I_{i^*}(t)I_{i^*}(t')I_{k^*}(t')) + \mathbb{E} (I_{i^*}(t)I_{k^*}(t)I_{k^*}(t')) \right] \sum_{k \in U_h} \tilde{X}_k^2(t)\tilde{X}_k^2(t'). \end{aligned} \quad (2)$$

Next, it can be shown that

$$\begin{aligned} A_3(t, t') = & \left[\mathbb{E} (I_{i^*}(t)I_{j^*}(t)I_{k^*}(t')) + \mathbb{E} (I_{i^*}(t)I_{k^*}(t')I_{j^*}(t')) \right] \\ & \times \left[- (N_h - 1)^2 \gamma_h(t, t) \gamma_h(t', t') + 2 \sum_{k \in U_h} \tilde{X}_k^2(t)\tilde{X}_k^2(t') \right] \\ & + 4 \mathbb{E} (I_{i^*}(t)I_{i^*}(t')I_{j^*}(t)I_{k^*}(t')) \left[- (N_h - 1)^2 \gamma_h^2(t, t') + 2 \sum_{k \in U_h} \tilde{X}_k^2(t)\tilde{X}_k^2(t') \right]. \end{aligned} \quad (3)$$

To compute $A_4(t, t')$, use the decomposition

$$\sum_{i,j,k,l \in U_h} \tilde{X}_i(t) \tilde{X}_j(t) \tilde{X}_k(t') \tilde{X}_l(t') = \sum_{\ell=1}^4 \sum_{\substack{i,j,k,l \in U_h \\ \mathcal{C}_{ijkl}=\ell}} \tilde{X}_i(t) \tilde{X}_j(t) \tilde{X}_k(t') \tilde{X}_l(t')$$

together with (1)–(3) to obtain

$$A_4(t, t') = \mathbb{E} \left(I_{i^*}(t) I_{j^*}(t) I_{k^*}(t') I_{l^*}(t') \right) \times \left[(N_h - 1)^2 \gamma_h(t, t) \gamma_h(t', t') + 2 (N_h - 1)^2 \gamma_h^2(t, t') - 6 \sum_{k \in U_h} \tilde{X}_k^2(t) \tilde{X}_k^2(t') \right]. \quad (4)$$

The proof is completed by gathering (1)–(4) and observing that all terms involving $\sum_{k \in U_h} \tilde{X}_k^2(t) \tilde{X}_k^2(t')$ are of negligible order $\mathcal{O}(N_h)$ thanks to (A1). \square

Proof of Proposition 3

Let k, l be two distinct units in a stratum U_h . Consider the Markov chain $\{(I_k + I_l)(\tau_r), r = 0, \dots, m\}$ which counts how many units among k, l are present in the sample at the successive replacement times. This chain has three possible states: 0, 1, and 2. For $r = 1, \dots, m$, the transition probability matrix

$$\mathbf{P}_r = \left(\mathbb{P} \left((I_k + I_l)(\tau_r) = j - 1 \mid (I_k + I_l)(\tau_{r-1}) = i - 1 \right) \right)_{1 \leq i, j \leq 3}$$

can be represented as

$$\mathbf{P}_r = \mathbf{P}_r^* + \mathbf{E}_r, \quad (5)$$

where

$$\mathbf{P}_r^* = \begin{pmatrix} (1 - \beta_r)^2 & 2(1 - \beta_r)\beta_r & \beta_r^2 \\ \alpha_h(1 - \beta_r) & \alpha_h\beta_r + (1 - \alpha_h)(1 - \beta_r) & (1 - \alpha_h)\beta_r \\ \alpha_h^2 & 2(1 - \alpha_h)\alpha_h & (1 - \alpha_h)^2 \end{pmatrix}$$

and $\beta_r = \mathbb{P}(k \in s_h(\tau_r) \mid k \notin s_h(\tau_{r-1})) = (f_h(\tau_r) - (1 - \alpha_h)f_h(\tau_{r-1})) / (1 - f_h(\tau_{r-1}))$.

Recall that $\alpha_h = \mathbb{P}(k \notin s_h(\tau_r) \mid k \in s_h(\tau_{r-1}))$. The matrix \mathbf{E}_r , whose cumbersome expression is not given here, is asymptotically negligible in comparison to \mathbf{P}_r^* . More precisely, (A5) guarantees that $\max_{r=1, \dots, m} \|\mathbf{E}_r\| = \mathcal{O}(1/N_h)$ as $N \rightarrow \infty$, where $\|\cdot\|$ denotes an arbitrary matrix norm. For simplicity we use the spectral norm $\|\mathbf{A}\| = \sup_{\mathbf{x} \neq 0} (\mathbf{x}' \mathbf{A}' \mathbf{A} \mathbf{x} / \mathbf{x}' \mathbf{x})^{1/2}$ henceforth.

The transition probability matrices \mathbf{P}_r have unit spectral norm. Using the binomial formula, the triangle inequality, and the inequality $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$ holding for all compatible matrices \mathbf{A} and \mathbf{B} , it follows that

$$\begin{aligned} \left\| \prod_{r=\nu(t)+1}^{\nu(t')} \mathbf{P}_r - \prod_{r=\nu(t)+1}^{\nu(t')} \mathbf{P}_r^* \right\| &\leq \sum_{r=1}^{\nu(t')-\nu(t)-1} \binom{\nu(t')-\nu(t)-1}{r} \left(\max_{q=\nu(t)+1, \dots, \nu(t')} \|\mathbf{E}_q\| \right)^r \\ &\leq \left(1 + \max_{r=1, \dots, m} \|\mathbf{E}_r\| \right)^m - 1 \\ &= \mathcal{O} \left(m \max_{r=1, \dots, m} \|\mathbf{E}_r\| \right) \end{aligned} \quad (6)$$

uniformly in $0 \leq t \leq t' \leq T$. Combining the previous results and condition (A5), it comes that

$$\prod_{r=\nu(t)+1}^{\nu(t')} \mathbf{P}_r = (1 + o(1)) \prod_{r=\nu(t)+1}^{\nu(t')} \mathbf{P}_r^*. \quad (7)$$

We now study the simpler product $\prod_{r=\nu(t)+1}^{\nu(t')} \mathbf{P}_r^*$. Without loss of generality, set $\nu(t) = 0$ and $\nu(t') = r$. Define $\mathbf{Q}_r = \prod_{l=1}^r \mathbf{P}_l^*$ and write $[\mathbf{A}]_{ij}$ for the (i, j) th coefficient of a matrix \mathbf{A} . It remains to compute $[\mathbf{Q}_r]_{13}$, $[\mathbf{Q}_r]_{23}$, and $[\mathbf{Q}_r]_{33}$.

Using the facts that \mathbf{Q}_r is a transition probability matrix and that $[\mathbf{P}_r^*]_{2j} = [\mathbf{P}_r^*]_{1j}^{1/2} [\mathbf{P}_r^*]_{3j}^{1/2}$ for $j = 1, 3$, it can be shown by induction that $[\mathbf{Q}_r]_{11}^{1/2} + [\mathbf{Q}_r]_{13}^{1/2} = 1$ for $r = 1, \dots, m$. Consequently $[\mathbf{Q}_r]_{13}^{1/2} = (1 - \beta_r - \alpha_h) [\mathbf{Q}_{r-1}]_{13}^{1/2} + \beta_r$, which can be reformulated as

$$[\mathbf{Q}_r]_{13}^{1/2} - f_h(\tau_r) = \frac{1 - \alpha_h - f_h(\tau_r)}{1 - f_h(\tau_{r-1})} \left([\mathbf{Q}_{r-1}]_{13}^{1/2} - f_h(\tau_{r-1}) \right). \quad (8)$$

Noting that $[\mathbf{Q}_1]_{13}^{1/2} - f_h(\tau_1) = -f_h(\tau_0)(1 - \alpha_h - f_h(\tau_1))/(1 - f_h(\tau_0))$ and iterating (8), we obtain the identity $[\mathbf{Q}_r]_{13}^{1/2} = f_h(\tau_r) - f_h(\tau_0) \lambda_h(\tau_0, \tau_r)$.

The arguments used to determine $[\mathbf{Q}_r]_{13}^{1/2}$ can be identically applied to $[\mathbf{Q}_r]_{33}^{1/2}$. Omitting the calculations, we directly state that $[\mathbf{Q}_r]_{33}^{1/2} = (1 - f_h(\tau_0)) \lambda_h(\tau_0, \tau_r) + f_h(\tau_r)$. The expressions of $[\mathbf{Q}_r]_{13}$ and $[\mathbf{Q}_r]_{33}$ can be checked by induction.

Finally we turn to $[\mathbf{Q}_r]_{23}$. The total probability formula yields

$$\begin{aligned} \mathbb{P}(k, l \in s(\tau_r)) &= \mathbb{P}(k, l \in s(\tau_r) | k, l \in s(0)) \mathbb{P}(k, l \in s(0)) \\ &\quad + 2 \mathbb{P}(k, l \in s(\tau_r) | k \in s(0), l \notin s(0)) \mathbb{P}(k \in s(0), l \notin s(0)) \\ &\quad + \mathbb{P}(k, l \in s_h(t') | k, l \notin s(0)) \mathbb{P}(k, l \notin s(0)). \end{aligned} \quad (9)$$

In view of (7) we obtain the approximation

$$f_h(\tau_r)^2 \sim \left(f_h(\tau_0)^2 [\mathbf{Q}_r]_{11} + 2 f_h(\tau_0) (1 - f_h(\tau_0)) [\mathbf{Q}_r]_{21} + (1 - f_h(\tau_0))^2 [\mathbf{Q}_r]_{31} \right). \quad (10)$$

The proof is completed by plugging the expressions of $[\mathbf{Q}_r]_{13}$ and $[\mathbf{Q}_r]_{33}$ in the previous relation. \square

Proof of Theorem 3

We start by finding the asymptotic expressions of $C_1(t, t')$ and $C_2(t, t')$ in Proposition 2. Exploiting (A3), the basic properties of SRSWOR and the independence of the $s_h(\tau_r)$, $r = 1, \dots, m$, under the full replacement design, it comes that

$$\begin{cases} C_1(t, t') \sim f_h(t) f_h(t') (1 - f_h(t)) (1 - f_h(t')) \\ C_2(t, t') \sim 2 f_h(t) f_h(t') (1 - f_h(t)) (1 - f_h(t')) \delta_{\nu(t)\nu(t')} \end{cases}$$

uniformly in $t, t' \in [0, T]$ as $N \rightarrow \infty$. Hence the second term in the right-hand side of (5.3) cancels out with the term in $C_1(t, t')$ of Proposition 2.

Writing $\text{Var}(\text{ISE}) = (2/N^2) \iint_{[0, T]^2} \phi_N(t, t') dt dt'$, we deduce that

$$\phi_N(t, t') \sim \left(\sum_{h=1}^H \frac{N_h}{N} \frac{1 - f_h(t)}{f_h(t')} \delta_{\nu(t)\nu(t')} \gamma_h(t, t') \right)^2 \quad (11)$$

uniformly in $t, t' \in [0, T]$ as $N \rightarrow \infty$. Hence, using (A2), (A4) and the mean value theorem, we get

$$\begin{aligned} \text{Var}(\text{ISE}) &\sim \frac{2}{N^2} \sum_{h, h'} \frac{N_h N_{h'}}{N^2} \sum_{r=1}^{m+1} \frac{(1 - f_h(\tau_r)) (1 - f_{h'}(\tau_r))}{f_h(\tau_r) f_{h'}(\tau_r)} \iint_{[\tau_{r-1}, \tau_r]^2} \gamma_h(t, t') \gamma_{h'}(t, t') dt dt' \\ &\sim \frac{2}{N^2} \sum_{h, h'} \frac{N_h N_{h'}}{N^2} \sum_{r=1}^{m+1} \frac{(1 - f_h(\tau_r)) (1 - f_{h'}(\tau_r))}{f_h(\tau_r) f_{h'}(\tau_r)} (\tau_r - \tau_{r-1})^2 \gamma_h(\tau_r, \tau_r) \gamma_{h'}(\tau_r, \tau_r) \\ &\sim \frac{2}{N^2} \sum_{h, h'} \frac{N_h N_{h'}}{N^2} \sum_{r=1}^{m+1} \frac{(1 - f_h(\tau_r)) (1 - f_{h'}(\tau_r))}{f_h(\tau_r) f_{h'}(\tau_r)} \frac{g(\tau_r)^2}{m^2} \gamma_h(\tau_r, \tau_r) \gamma_{h'}(\tau_r, \tau_r) \\ &\sim \frac{2}{m N^2} \sum_{h, h'} \frac{N_h N_{h'}}{N^2} \int_0^T \frac{(1 - f_h(t)) (1 - f_{h'}(t))}{f_h(t) f_{h'}(t)} g(t)^2 \gamma_h(t, t) \gamma_{h'}(t, t) dt. \quad \square \end{aligned}$$

Proof of Theorem 4

This result is established along the same lines as Theorem 3. We start by finding the asymptotic expression of $C_1(t, t')$ and $C_2(t, t')$ in Proposition 2 under partial replacement. In view of Lemmas 1-2 and Proposition 3, it holds that for all $0 \leq t \leq t' \leq T$, as $N \rightarrow \infty$,

$$\begin{aligned}
& \mathbb{E} (I_{i^*}(t) I_{k^*}(t')) \\
&= \mathbb{P} (k^* \in s(t') | i^*, k^* \in s(t)) \mathbb{P} (i^*, k^* \in s(t)) + \mathbb{P} (k^* \in s(t') | i^* \in s(t), k^* \notin s(t)) \mathbb{P} (i^* \in s(t), k^* \notin s(t)) \\
&= \mathbb{P} (k^* \in s(t') | k^* \in s(t)) \mathbb{P} (i^*, k^* \in s(t)) + \mathbb{P} (k^* \in s(t') | k^* \notin s(t)) \mathbb{P} (i^* \in s(t), k^* \notin s(t)) \\
&\sim [(1 - f_h(t)) \lambda_h(t, t') + f_h(t')] f_h^2(t) + [f_h(t') - f_h(t) \lambda_h(t, t')] f_h(t) (1 - f_h(t)) \\
&= f_h(t) f_h(t').
\end{aligned}$$

By symmetry this expression holds for all $t, t' \in [0, T]$. Similarly, we find that

$$\left\{ \begin{array}{ll} \mathbb{E} (I_{i^*}(t) I_{j^*}(t) I_{k^*}(t')) & \sim f_h^2(t) f_h(t'), \\ \mathbb{E} (I_{i^*}(t) I_{k^*}(t') I_{l^*}(t')) & \sim f_h(t) f_h^2(t'), \\ \mathbb{E} (I_{i^*}(t) I_{j^*}(t) I_{k^*}(t') I_{l^*}(t')) & \sim f_h^2(t) f_h^2(t'), \\ \mathbb{E} (I_{i^*}(t) I_{i^*}(t') I_{k^*}(t) I_{k^*}(t')) & \sim [(1 - f_h(t)) \lambda_h(t, t') + f_h(t')]^2 f_h^2(t), \\ \mathbb{E} (I_{i^*}(t) I_{i^*}(t') I_{j^*}(t) I_{k^*}(t')) & \sim [(1 - f_h(t)) \lambda_h(t, t') + f_h(t')] f_h^2(t) f_h(t'). \end{array} \right.$$

Therefore

$$\left\{ \begin{array}{l} C_1(t, t') \sim f_h(t) f_h(t') (1 - f_h(t)) (1 - f_h(t')) \\ C_2(t, t') \sim 2 f_h^2(t) (1 - f_h(t))^2 \lambda_h^2(t, t') \end{array} \right. .$$

Combining the previous result and Proposition 2, it stems from (5.3) that

$$\sum_{i,j,k,l \in U_h} \Delta_{ijkl}(t, t') X_i(t) X_j(t) X_k(t') X_l(t') \sim 2 f_h^2(t) (1 - f_h(t))^2 \lambda_h^2(t, t') \gamma_h^2(t, t') N_h^2. \quad (12)$$

On the other hand the inter-strata contribution to (5.2) can be simplified thanks to Theorem 2:

$$\sum_{i,k \in U_h} \frac{\Delta_{ik}(t, t')}{f_h(t) f_h(t')} X_i(t) X_k(t') = N_h \frac{1 - f_h(t)}{f_h(t')} \gamma_h(t, t') \lambda_h(t, t'). \quad (13)$$

With the notations of the proof of Theorem 3, one deduces from (12) and (13) that

$$\phi_N(t, t') \sim \left(\sum_{h=1}^H \frac{N_h}{N} \frac{1 - f_h(t)}{f_h(t')} \gamma_h(t, t') \lambda_h(t, t') \right)^2 \quad (14)$$

for all $t, t' \in [0, T]$, as $N \rightarrow \infty$. To apply the dominated convergence theorem, it suffices to check that the ϕ_N , $N \geq 1$, are uniformly bounded on $[0, T]^2$. In view of (A1) and (A2), the right-handside of (14) has a finite number of terms, the terms $(1 - f_h(t))/f_h(t')$ and $\gamma_h(t, t')$ are uniformly bounded with respect to h , $t, t' \in [0, T]$ and N , and $|\lambda_h(t, t')| \leq 1$ as a product of eigenvalues of transition probability matrices. The dominated convergence theorem thus applies, which concludes the proof of Theorem 4. \square

Proof of Corollary 1

In view of Theorem 4, it suffices to show that $\lambda_h(t, t') \sim \exp(-c_h |G(t) - G(t')|)$ as $N \rightarrow \infty$. By condition (i) of the corollary, we first see that $\lambda_h(t, t')$ is equal to $(1 - \alpha_h/(1 - f_h))^{\nu(t) - \nu(t')}$. By condition (A2), the term $|\nu(t) - \nu(t')|$ works out as $(m + 1) |G(\tau_{\nu(t)}) - G(\tau_{\nu(t')})|$, which in turn is asymptotic to $m |G(t) - G(t')|$. Finally exploiting condition (ii) of the corollary and the approximation $\ln(1 - x) \sim (-x)$ as $x \rightarrow 0$, we obtain the sought equivalent for $\lambda_h(t, t')$. \square

References

- Cardot, H., Chaouch, M., Goga, C., and Labruère, C. (2010). Properties of design-based functional principal components analysis. *J. Statist. Plann. Inference*, **140**, 75-91.
- Cardot, H., Degras, D., and Josserand, E. (2012). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *Bernoulli*. Forthcoming, also available at <http://arxiv.org/abs/1105.2135>.
- Cardot, H. and Josserand, E. (2011). Horvitz-Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, **98**, 107-118.
- Chiky, R., Cubillé, J., Dessertaine, A., Hébrail, G., and Picard, M.-L. (2008). Échantillonnage spatio-temporel de flux de données distribués. In Guillet, F. and Trousse, B., editors, *EGC'08*, pages 169–180.
- Degras, D. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statist. Sinica*, **21**, 1735-1765.
- Eckler, A. R. (1955). Rotation sampling. *Ann. Math. Statist.*, **26**, 664–685.

- Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York. Theory and practice.
- Fuller, W. A. (2009). *Sampling statistics*. Wiley Series in Survey Methodology. Hoboken, NJ: John Wiley & Sons.
- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663-685.
- Marbini, A. D. and Sacks, L. E. (2003). Adaptive sampling mechanisms in sensor networks. In *London Communications Symposium*, London, UK.
- Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **12**, 241-255.
- Rao, J. N. K. and Graham, J. E. (1964). Rotation designs for sampling on repeated occasions. *J. Amer. Statist. Assoc.*, **59**, 492-509.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer Series in Statistics. Springer, New York.
- Tikkiwal, G. and Gupta, A. (1991). Estimation of population mean under successive sampling scheme when various weights and regression coefficients are unknown. *Biom. J.*, **33**, 529-538.
- Wolter, K. M. (1979). Composite estimation in finite populations. *J. Amer. Statist. Assoc.*, **74**, **367**, 604-613.

DePaul University

E-mail: ddegrasv@depaul.edu